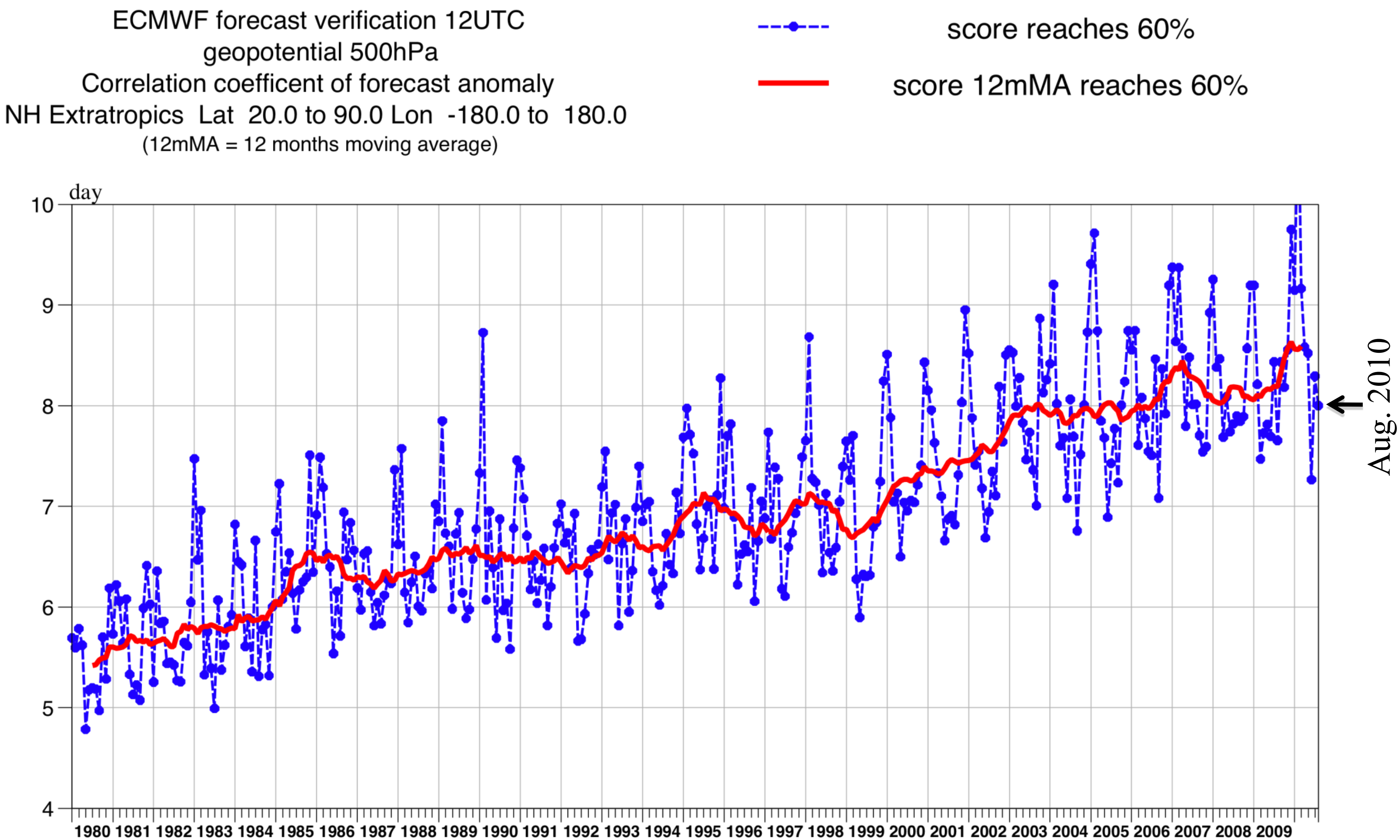
NWP Verification: what can be learnt?

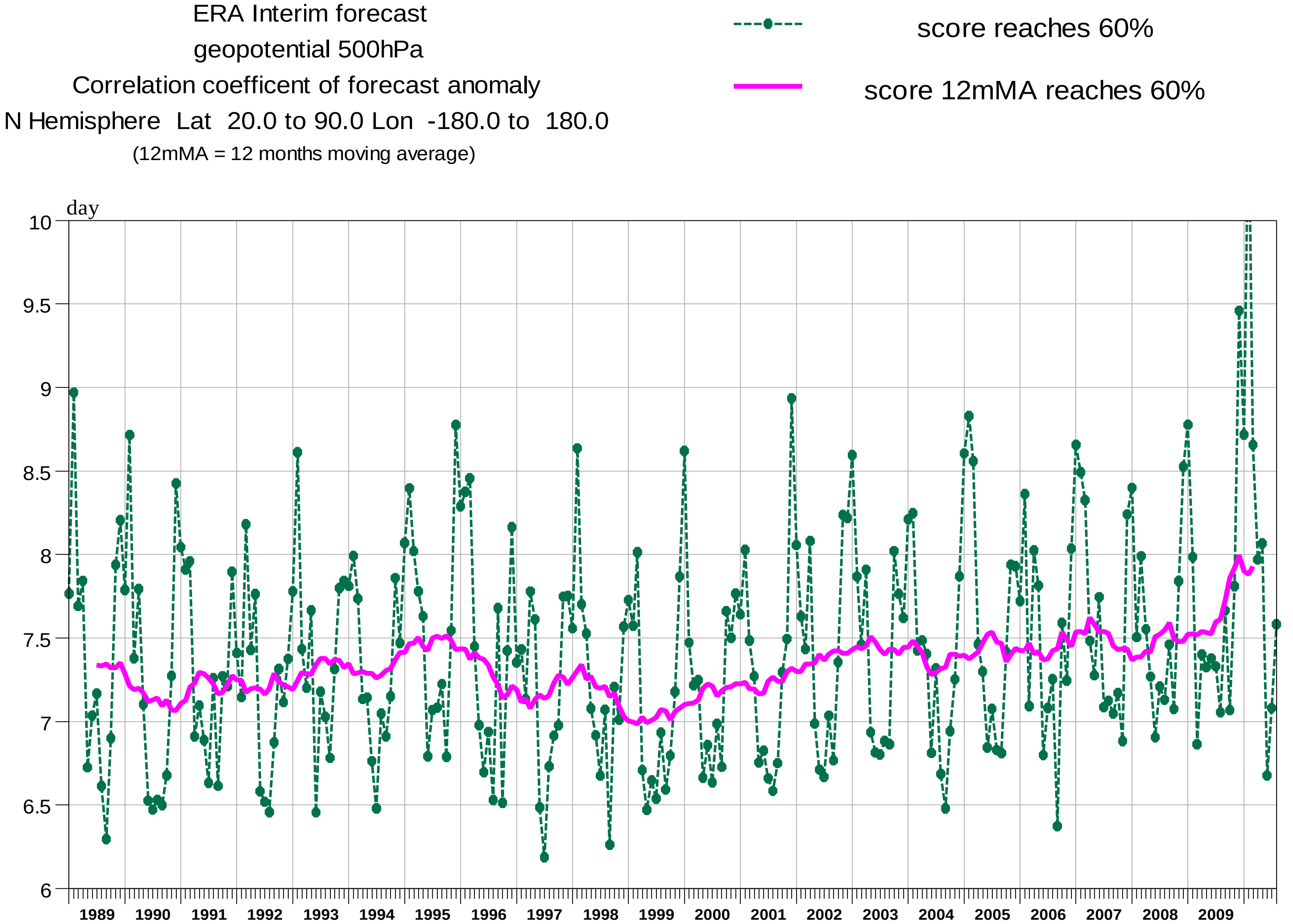
Anna Ghelli
ECMWF, Reading, UK
anna.ghelli@ecmwf.int

ECMWF

Time series Acc=60% N hemisphere



Time series Acc=60% N hemisphere ERA Interim forecast



ECMWF

OUTLINE

- Verification: WHY?
- Metrics used in NWP
- What is the truth?
    - Observations (what does the model produce?)
    - Analysis
- Spatial methods
- Suggestions on a verification framework

ECMWF

Why verify?

- Administrative purpose
  - Monitoring performance
- Scientific purpose
  - Identifying and correcting model flaws
  - Forecast improvement
- Economic purpose
  - Improved decision making
  - “Feeding” decision models or decision support systems
- Forecasters
  - Understanding biases
  - Understanding strengths and weaknesses of models

Verification

Forecast Attributes

Observations availability/analysis

Visualisation

Reference system

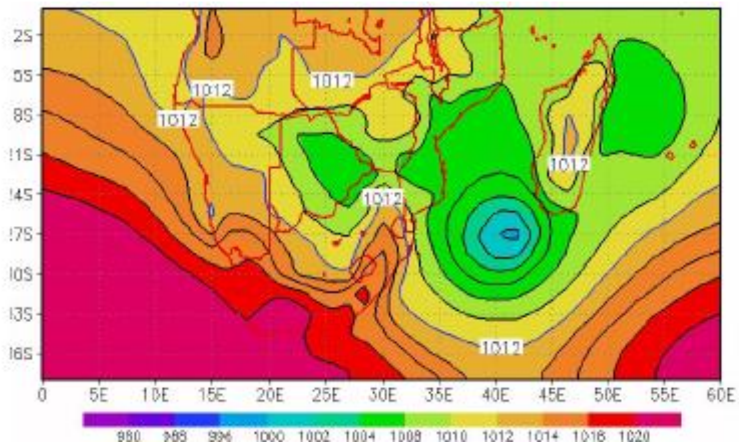
The questions:

- In what locations does the model have the best performance?
- Are there regimes in which the forecasts are better or worse?
- Is the probability forecast well calibrated (i.e., reliable)?
- Do the forecasts correctly capture the natural variability of the weather?
- Is the genesis in the right location?
- Is the landfall accurate?
- Which model is better according to some specified scoring rule?
- Is there any systematic bias?

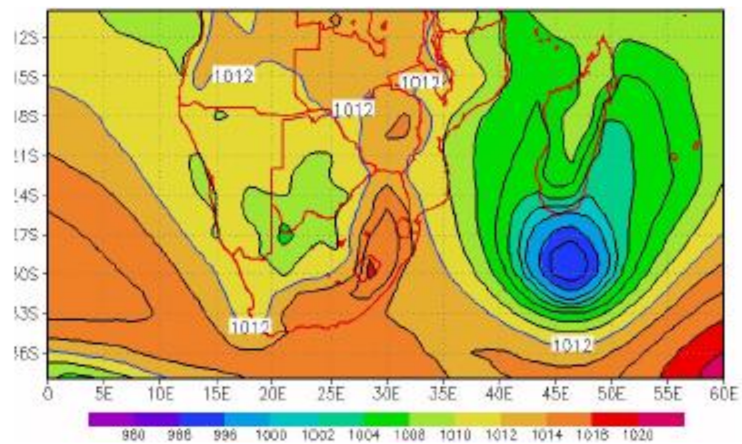
Which verification technique?

Tropical cyclone forecast

Observed



3-day forecast



Who are our users?
Aggregation/stratification?
What do we want to measure?

Bias
Position/ intensity error
Attribute of features
Reliability
Discrimination

ECMWF

Forecast quality versus forecast value

- A forecast has high **QUALITY** if it predicts the observed conditions well according to some objective or subjective criteria.

- A forecast has **VALUE** if it helps the user to make a better decision.

Quality but no value

Value but no quality

Scores: formulation

- *Root Mean Square Error:*

$$E = \sqrt{\overline{(fc - an)^2}}$$

- *Bias:*

$$BIAS = \overline{FC - OBS}$$

- *Mean Absolute Error :*

$$MAE = \overline{|FC - OBS|}$$

- *Anomaly Correlation:*

$$ACC = \frac{\overline{(fc - c)(an - c)}}{\sqrt{A_{fc} A_{an}}}$$

$$A_{fc} = \overline{(fc - c)^2}$$

$$A_{an} = \overline{(an - c)^2}$$

Measures accuracy
Range: 0 to infinity perfect score = 0

Measures bias
Range: -infinity to +infinity perfect score = 0

Measures accuracy
Range: 0 to infinity perfect score = 0

Measures accuracy
Range: -100% to 100% perfect score = 100%

Contingency tables

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	Hit	False alarm	Fc Yes
No	Miss	Correct non-event	Fc No
Marginal total	Obs Yes	Obs No	Sum total

Event forecast	Event observed		
	Yes	No	Marginal total
Yes	a	b	a + b
No	c	d	c + d
Marginal total	a + c	b + d	a + b + c + d =n

Frequency Bias

$$FBI = B = \frac{(a+b)}{(a+c)}$$

Hit Rate

$$H = POD = \frac{a}{(a+c)}$$

False Alarm Rate

$$= \frac{b}{(b+d)}$$

Equitable Threat Score

$$ETS = \frac{(a-a_r)}{(a+b+c-a_r)} \qquad a_r = \frac{(a+b)(a+c)}{n}$$

True Skill Score (also known as Pierce's Skill Score)

$$TSS = PSS = \frac{ad-bc}{(a+c)(b+d)}$$

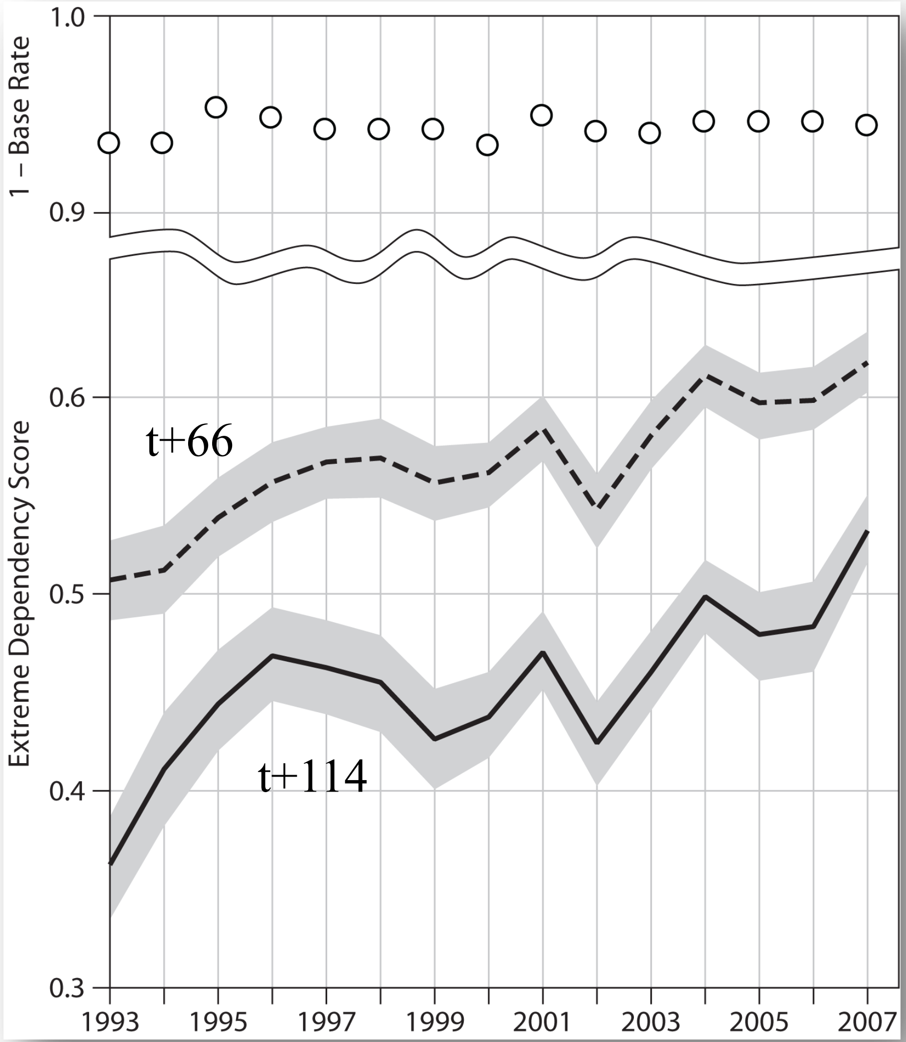
Evaluating model precipitation forecasts

Evaluating “rare” events

Rainy season in Europe
October- April

Shaded areas represent the 90% confidence interval

$$EDS = 2\frac{\ln\left(\frac{a+c}{n}\right)}{\ln\left(\frac{a}{n}\right)} - 1$$



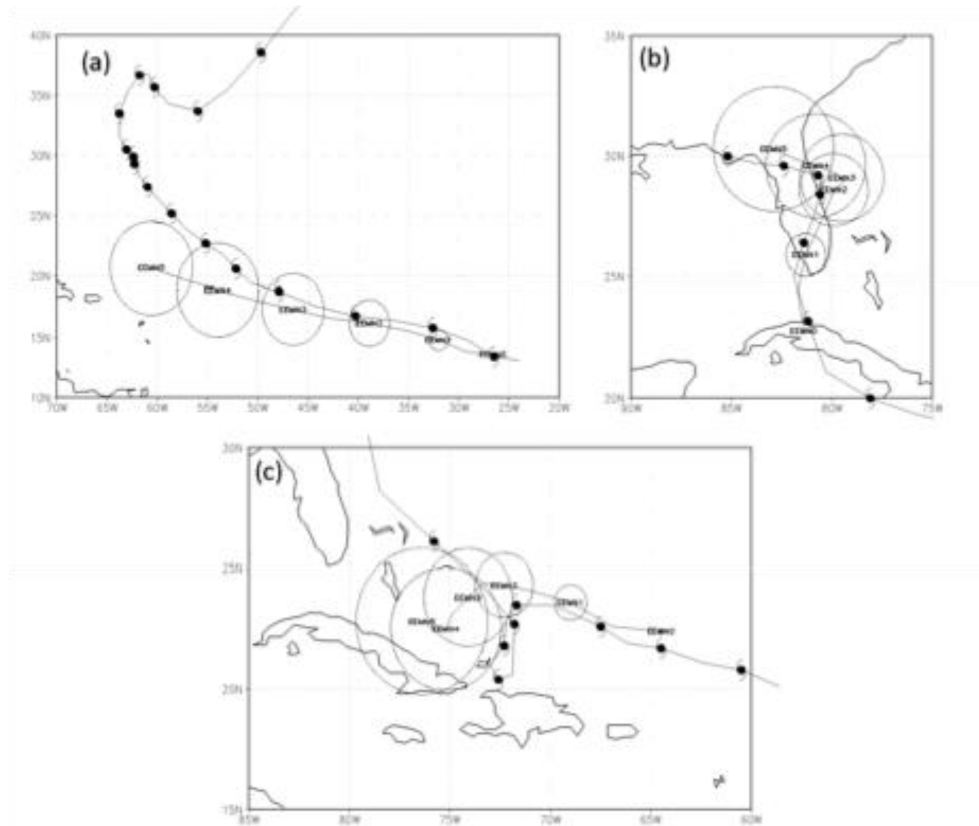
ECMWF

Observation/model matching

Identifying observations that represent the forecast event

Gridded forecasts and observations need to be matched

- Point-to-grid
  - Match obs to closest gridpoint
  - Average all observations within the grid box?

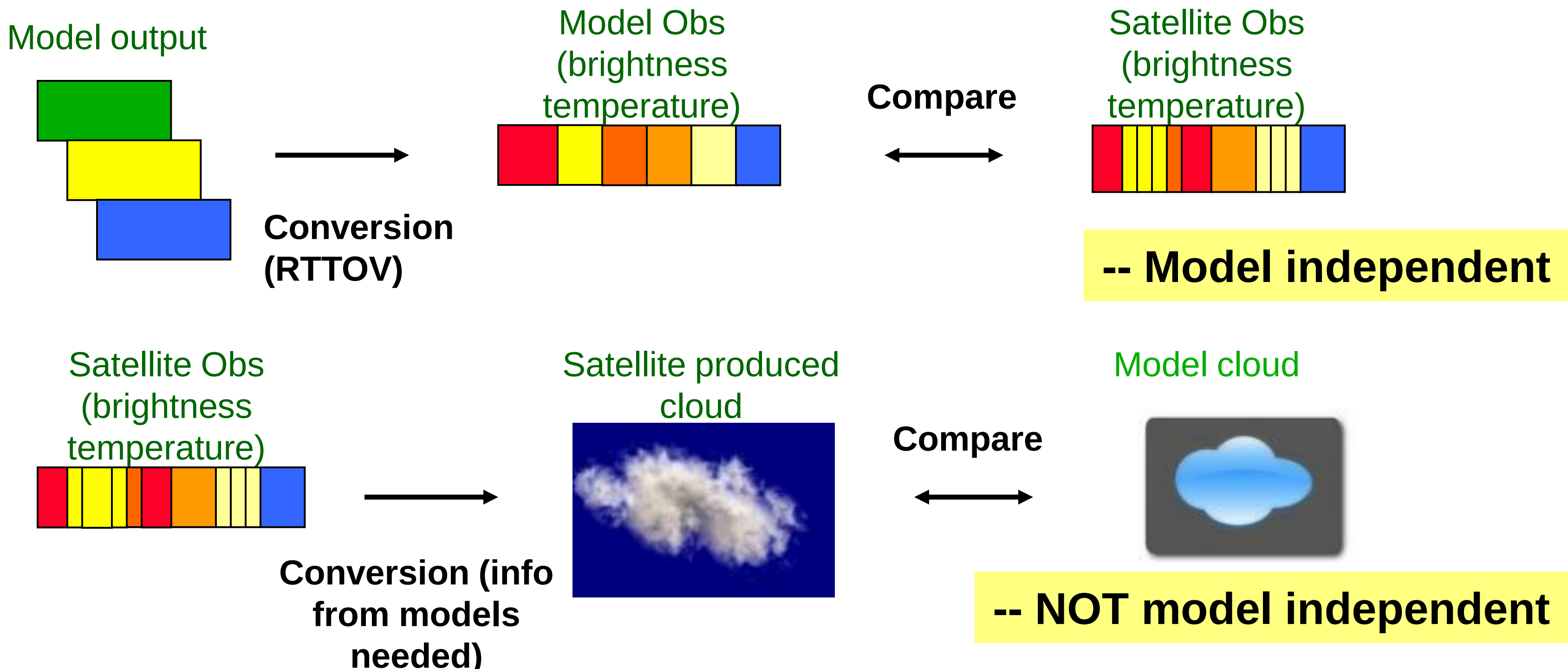


- Grid-to-point
  - Interpolate?
  - Take largest value?

The matching game:
Strive for an independent dataset

Approaches:

- Model to observations ➔ model output is manipulated to become comparable to observations
- Observations to model ➔ observations are manipulated to become comparable to model output

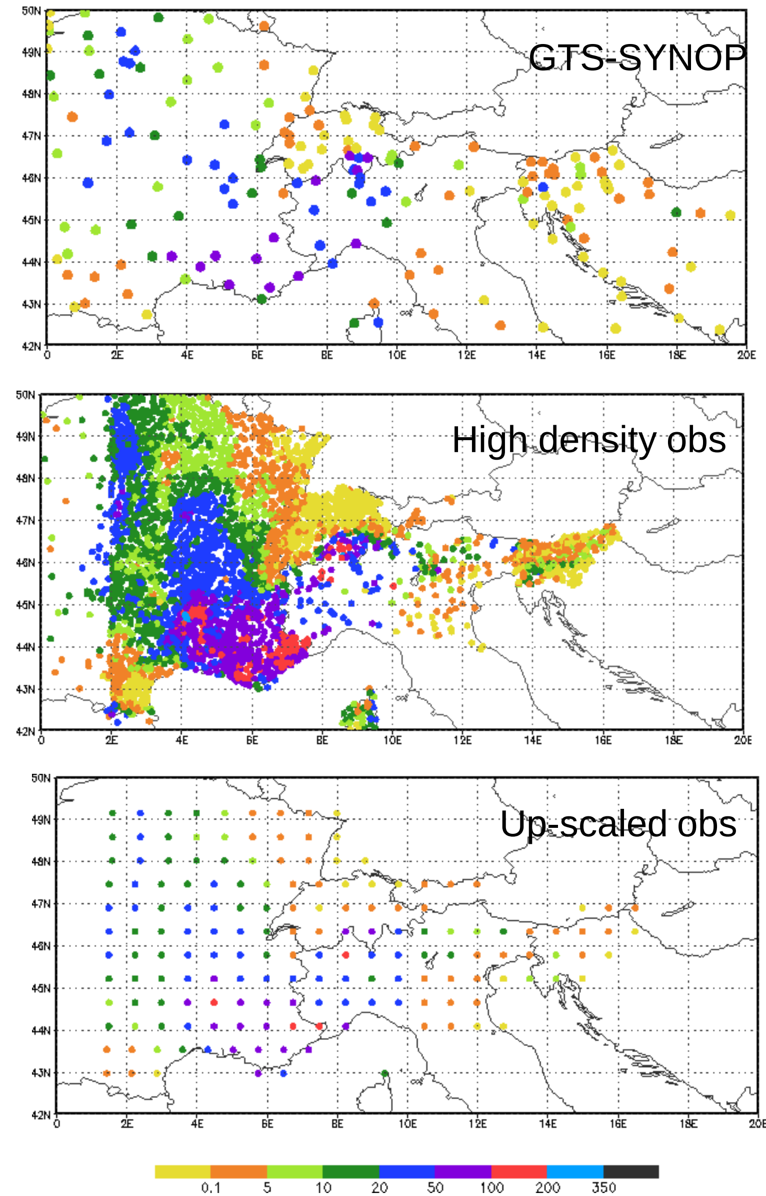


-- A different perspective --

- The model will not produce exact results for scales smaller than its own spatial scale.
- **Comparisons between model forecast value and observations rely on either interpolation or close neighbour method.**
- Precipitation shows large variability and the precipitation amount measured in a specific location, may not be representative of an area (under sampling)
- **Precipitation forecast should be interpreted as an areal value rather than a point value.**
- High resolution network stations used to produce mean values of precipitation to be attributed to each grid-point. Such values are then compared to the model forecast. “Up-scaling” of the information contained in the observations to make comparisons that are fairer to model

The Up-scaling technique

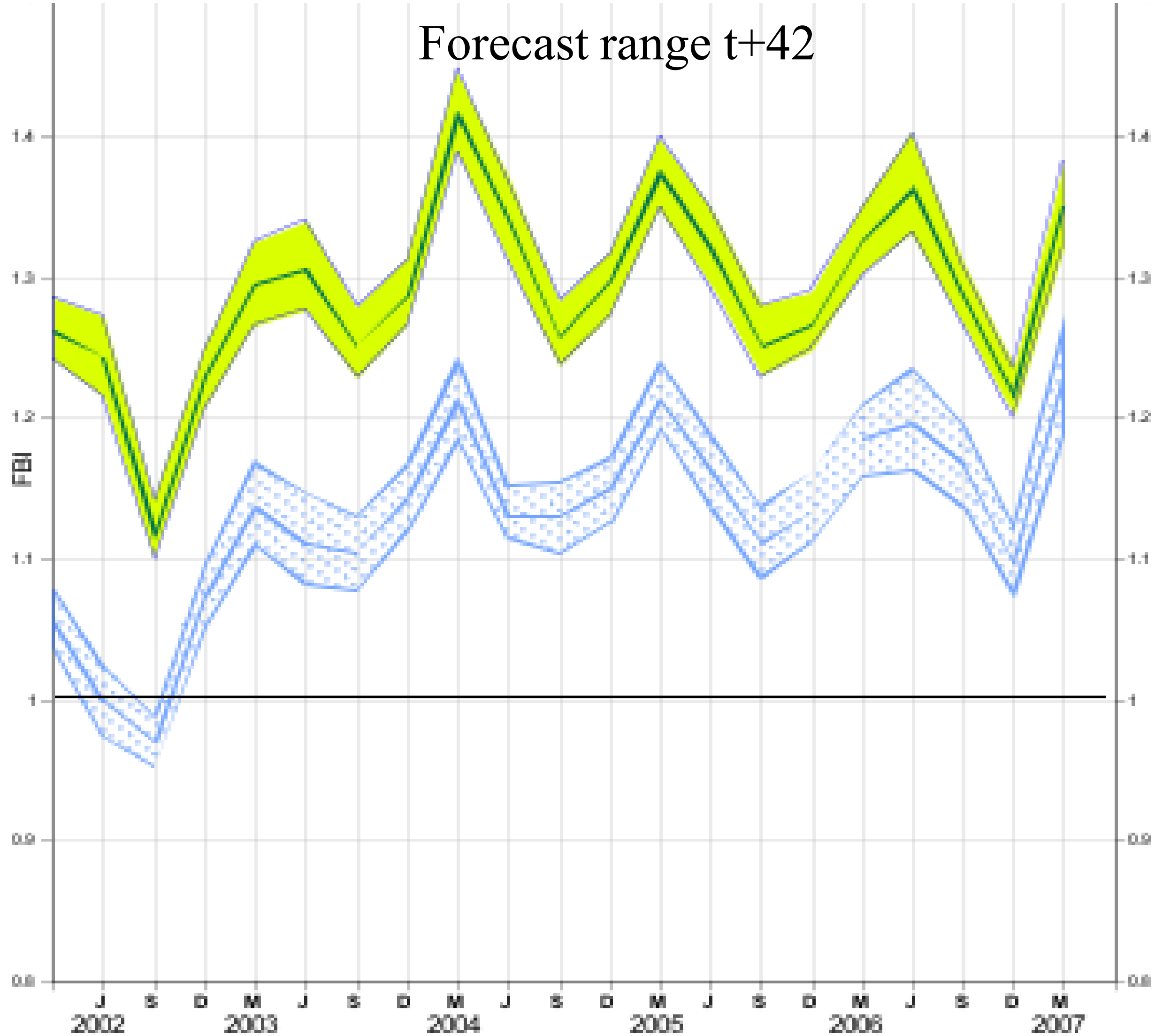
- There are many methods available to up-scale observations to the model resolution
- We have used a simple averaging procedure of all the observations contained in a model gridbox
- Alps: SYNOP coverage, high-density observations and up-scaled observed values for Sept. 20, 1999



-- A different perspective --

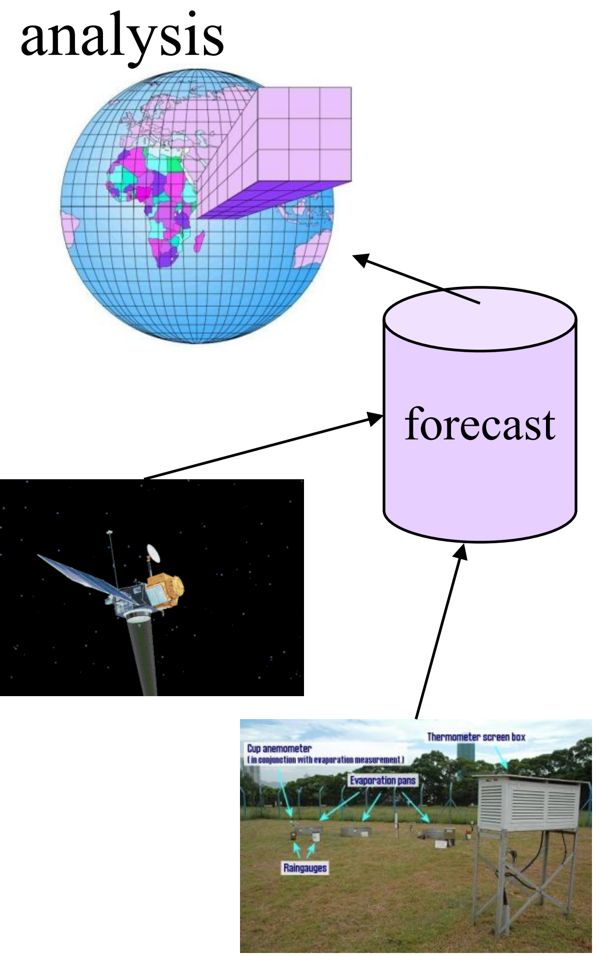
FBI
Threshold > 1mm/24h

Cyan: **precipitation analysis** (shaded area indicate uncertainty)
Green: **Synops on GTS** (shaded area indicates uncertainty)



ECMWF

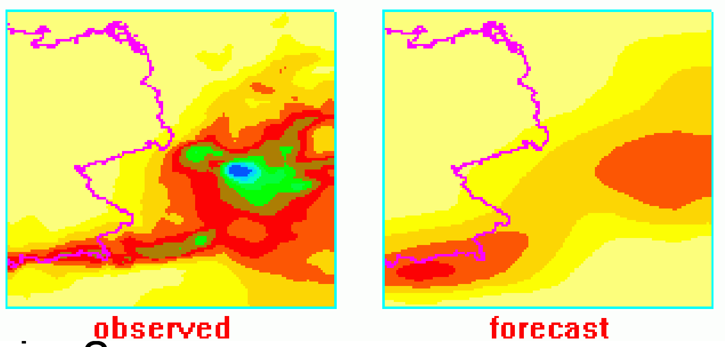
The role of the analysis in verification



- **Analyses are model dependent**
  - **Allows to use a number of different type of sensors to provide a coherent analysis for the model ➔ this out-weight the drawback of model contamination**
  - **Good if used for specific purposes e.g. when performance needs to be assessed for scales that the model can resolve and for comparison of same model (operational vs. experimental suite)**
  - **Multi-analysis against observations scores better than single analysis**
  - **Use of randomly drawn analyses for comparative verification of multiple models.**

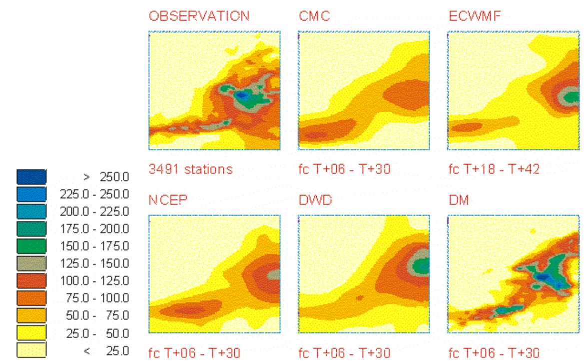
Scores: what they can/cannot offer

- Overall measures of skill (accuracy, bias)
- Minimal diagnostic information
- Cannot answer the following questions:
  - What went wrong? What was right?
  - Does the forecast look realistic?
  - How can I improve the forecast?
  - How can I use the forecast to make a decision?



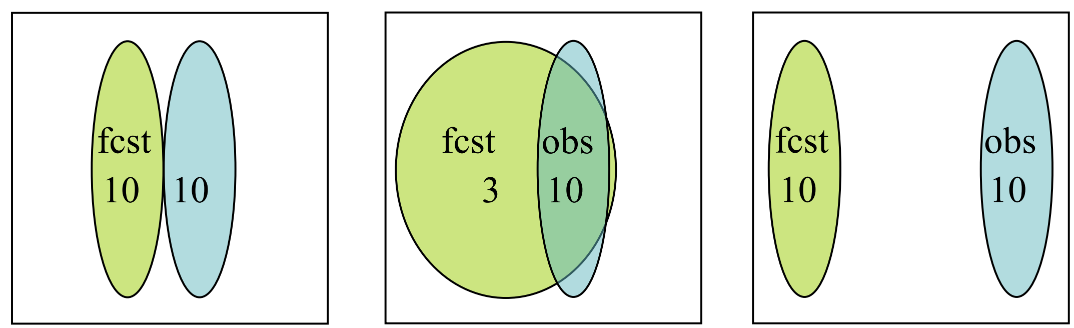
ECMWF

Spatial verification



Standard verification

- Need matching between forecast and observation
- Double penalty
- Do not say source of error
- Do not say how to improve forecasts



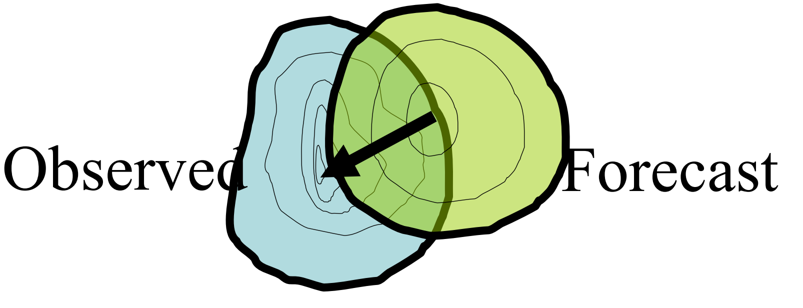
Hi res forecast
RMS ~ 4.7
POD=0, FAR=1
TS=0

Low res forecast
RMS ~ 2.7
POD~1, FAR~0.7
TS~0.3

Hi res forecast
RMS ~ 4.7
POD=0, FAR=1
TS=0

Feature-based approach (CRA)
Ebert and McBride, J. Hydrol., 2000

- **Define entities using thresholds (Continuous Rain Areas)**
- **Horizontally translate forecasts until a matching pattern criterion is met:**
  - **Minimum total squared error between forecast and observation**
  - **Maximum correlation**
  - **Maximum overlap**
- **The displacement is the vector difference between the original and the final location of the forecasts**



Feature based approach (CRA)

- **Total mean squared error**

$$MSE_{total} = MSE_{displacement} + MSE_{volume} + MSE_{pattern}$$

- The **displacement error** is the difference between the mean squared error before and after translation

$$MSE_{displacement} = MSE_{total} - MSE_{shifted}$$

- The **volume error** is the bias in mean intensity

$$MSE_{volume} = (\overline{F} - \overline{X})^2$$

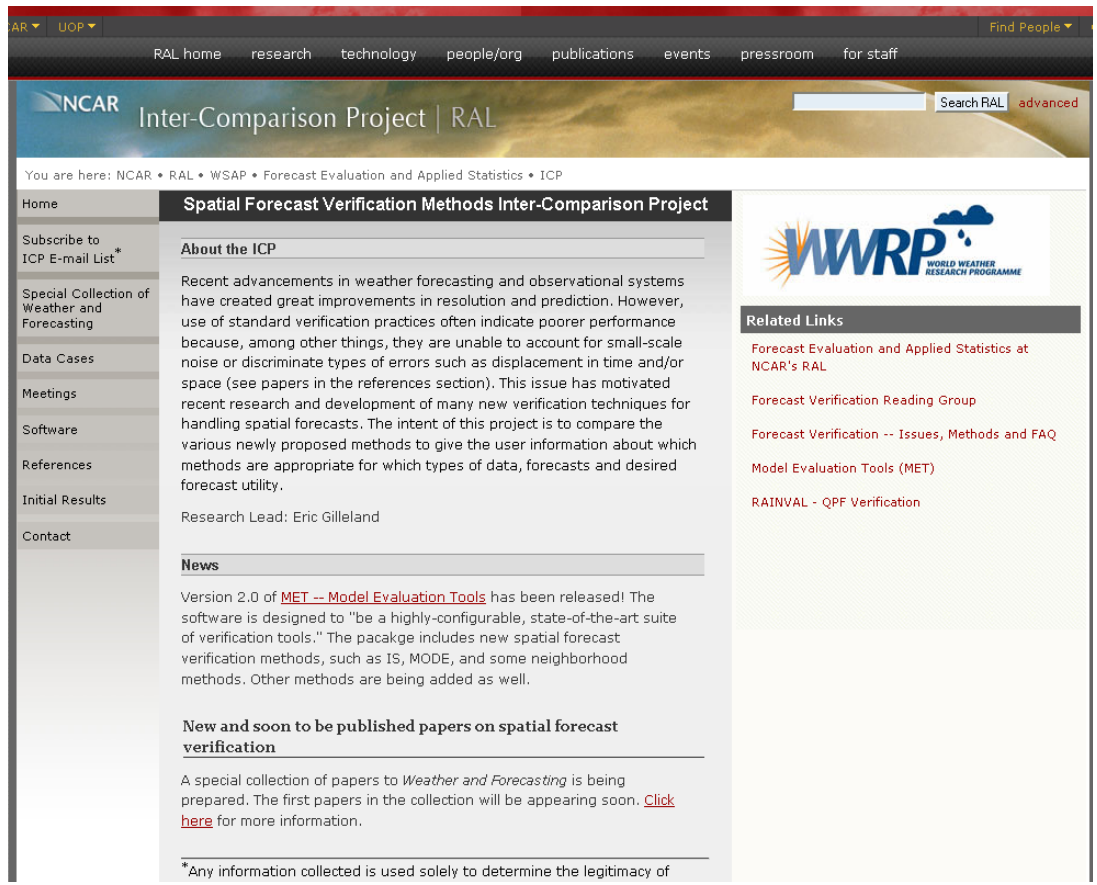
Where $\overline{F}$ and $\overline{X}$ are the mean forecast and observed values after shifting

- The **patter error**, computed as residual, accounts for the difference in structure

$$MSE_{pattern} = MSE_{shifted} - MSE_{volume}$$

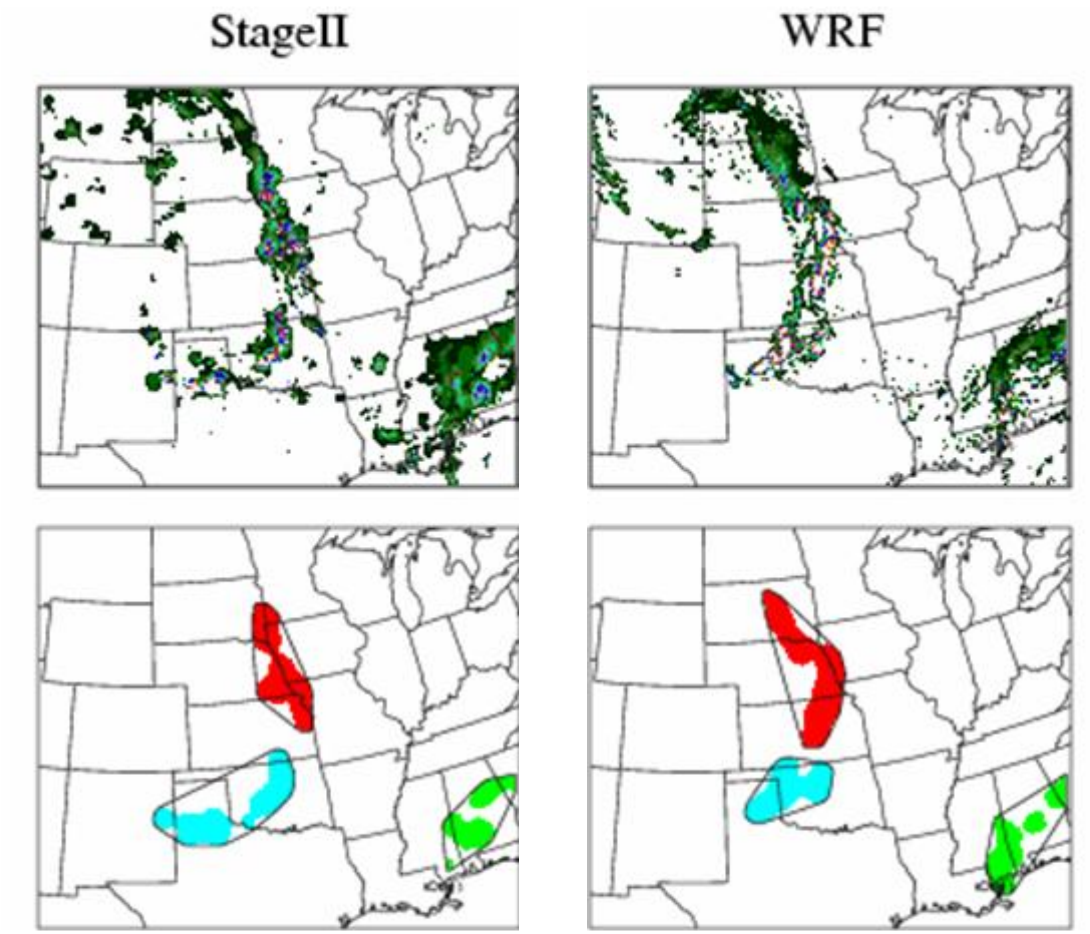
Spatial Verification Intercomparison Project

- http://www.ral.ucar.edu/projects/icp/index.html
  - Test cases
  - Results
  - Papers
  - Code



ECMWF

MODE object matching/merging



24h forecast of 1h rainfall on 1 June 2005

Compare attributes:

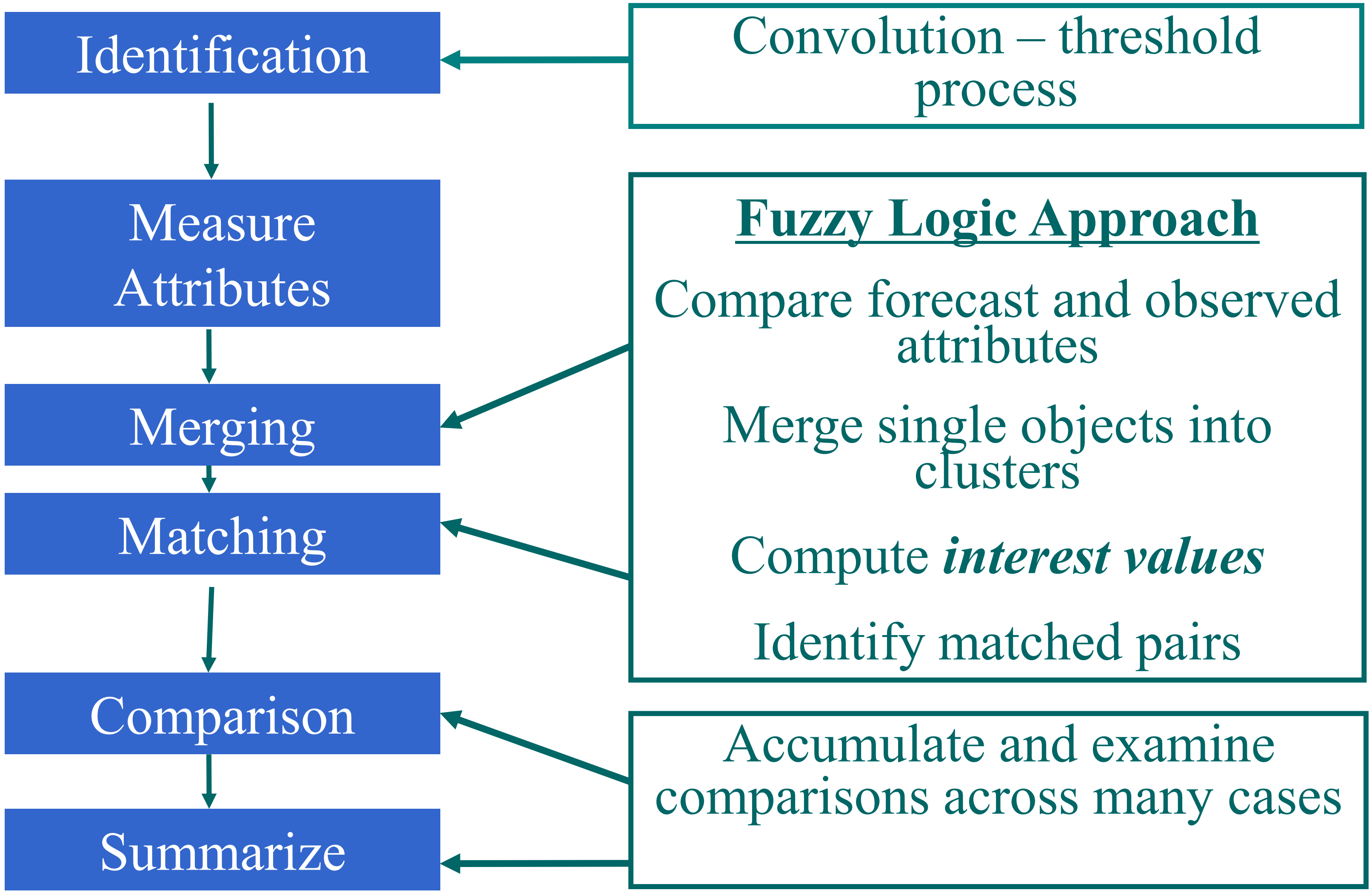
- - centroid location
- - intensity distribution
- - area
- - orientation
- - etc.

When objects not matched:

- - false alarms
- - missed events
- - rain volume
- - etc.

ECMWF

MODE methodology



ECMWF

Designing a verification framework

- **Establish who are the users of the verification**
  - **What is the real meaning of the parameter calculated in the model? It is an areal quantity or a point value? This may have some repercussions in the way the scores/data are calculated.**
- **Define a set of top level scores (administrative purposes/ economic purpose)**
- **Define a complementary set of scores which may address needs of specific users (scientific purpose/user purpose)**
  - **Time series will show trends, but case studies are relevant to understand what went wrong!**
  - **Use confidence intervals on the scores**